EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 37: April 21, 1995

Lecturer: Professor Ron Cole *Scribe: Su-Lin Wu*

In this lecture, Prof. Ron Cole from the Center for Spoken Language Understanding at the Oregon Graduate Institute discusses his experiences from two major speech recognition projects; Spoken Letter Recognition and a Voice Response Questionnaire for the U.S. Census.

Prof. Cole's original slides are available in the class archives.

37.1 Introduction

The Center for Spoken Language Understanding (CSLU) attacks core problems of speech recognition. The Center has students, who develop new algorithms and techniques, and staff, who implement the new technology in working systems. The Center plans to build systems for use by millions of people. Towards this end CSLU receives funding from many companies, including US West, who are interested in telephone-based systems.

37.2 Spoken Letter Recognition

Given a name spelled with pauses between letters by any speaker, this project's goal is to retrieve the correct name from a database of 50,000 names.

The project began with collecting speech samples over the telephone, a large and difficult task by itself.

The task of recognizing letters is not easy because many letters sound alike. For example, the letters in the "E" set (B D E P T G V Z C) all sound very similar. Continuous alphabet recognition is too hard because of the high confusibility of the letters. This task requires estimation of pitch or formant tracking, but no one has yet implemented this.

Instead, using the pauses the speaker supplies between letters, the utterance can be segmented and the letter boundaries located. Why is segmentation important? Segmentation helps recognition systems locate phoneme boundaries for feature extraction. For example, in "B" and "D", all the information for distinguishing between the two letters is in the beginning. With segmentation, the recognition system can focus on the beginning of the letter and use other techniques for fine discrimination.

37.2.1 System Details

The current system uses PLP to extract features from acoustic input. The features are averaged over seven regions surrounding the frame to be classified.

A neural network was trained to learn the relationship between features and acoustic inputs. The network is then expected to be able to generalize and classify new inputs. Training data was produced by a process of forced and unforced alignments, so that the network could be trained on background noise, as well as the letters.

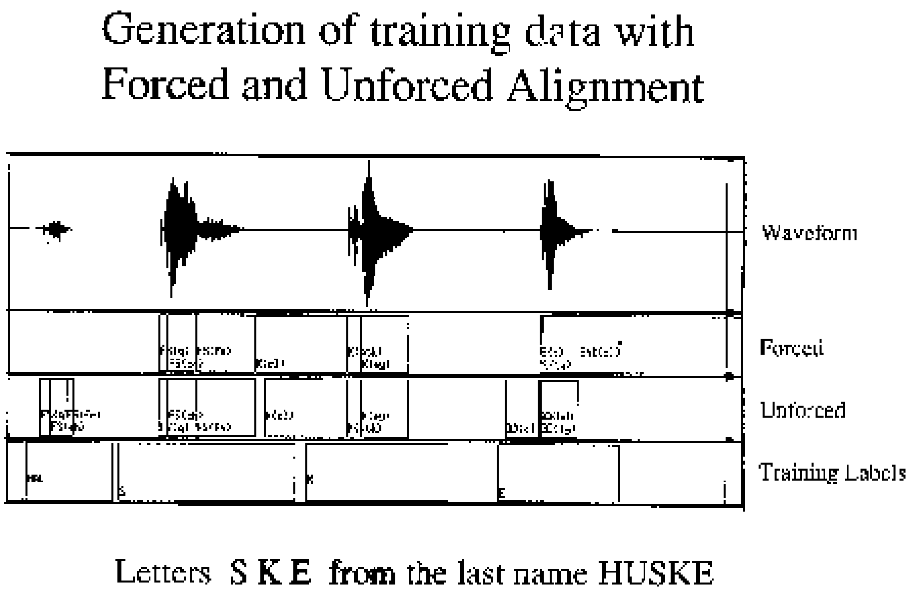

Figure 37.1: Forced and Unforced Alignment

Forced alignment means the recognition system is provided the correct sequence of sounds that occur in an utterance. With this information the system is very good at labelling the parts of the waveform with the appropriate sound. After training on these labels, an unforced alignment is performed and and the network labels the waveform as best it can. It often picks up background noise such as coughs, bangs and taps, and labels them as speech sounds. By comparing the labels generated from the forced and unforced alignment, the background noise can be identified and labeled "NAL" or Not A Letter. In this manner, training data for the background noise can be obtained (Figure 37.1).

The recognition system performs two passes; a first pass to segment the acoustic input and a second pass with a knowledge-based system to get the best answers. Initially, sounds are classified into broad categories of similar-sounding acoustic units and phonemes (Figure 37.2). Each letter has a simple pronunciation model where each state of the model corresponds to an output of the phonetic classifier. A Viterbi search finds the highest-scoring sequence of letters and provides phonetic boundaries.

The system then uses the output of the letter classifier in Figure 37.2 as a probability. Name models are stored as a tree and another search is used to find the best name, allowing for letter insertions and deletions.

37.2.2 Results

On a test with 300 spelled last names and a database of 50,000 last names, the results were that the system recognized the correct name as its top choice 90.7% of the time. The correct name was in the top two choices 94.0% of the time and in the top three 95.7% of the time.

On a task with a training set of 800 speakers (12,568 letters) and a test set of 400 speakers (4,214 letters) the system recognized the correct letter as its top choice 89.0% of the time, with .07% insertions. Names were recognized as the first choice 87.3% correctly with 1.5% insertions.

In a building with about 1000 people, this phone system would perform well in allowing callers to spell the names of people they wish to contact.

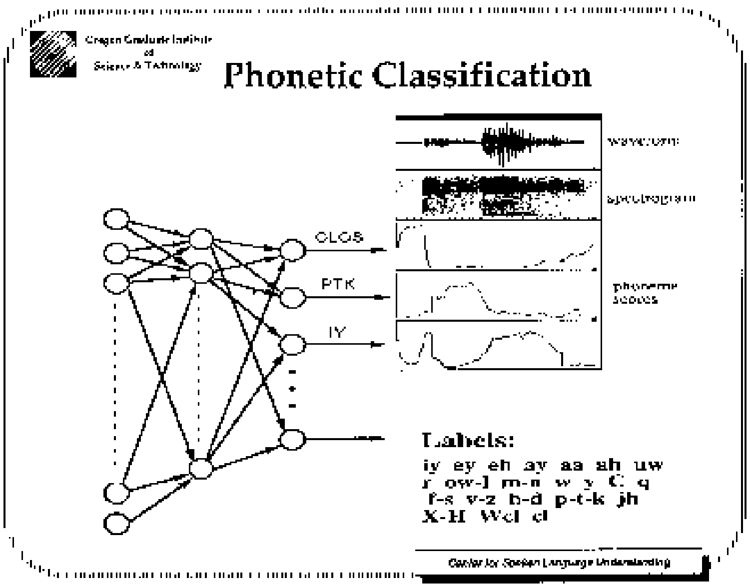

Figure 37.2: Phonetic Classification

37.3 A Prototype Voice-Response Questionnaire for the U.S. Census

The Census project had two phases. The first was to demonstrate the feasibility of a telephone-based questionnaire. For this CSLU created a question and answer dialogue and demonstrated its effectiveness. Then a prototype system was developed and tested. The second phase was to deploy a system at a Census pretest. This would incorporate a new and expanded protocol, and also the abilities to handle 20 simultaneous calls, detect and deal with barge-in (where the caller talks over the prompt), and bail out to a human operator. A new operator's graphical interface would facilitate human verification of calls. This system was installed in Jefferson, Indiana.

37.3.1 Background

Why automate the Census? Voice access would allow the census to more easily include illiterate people, non-English speakers and others who are disadvantaged. Compared to the current method with paper and people to follow up, a voice system would be cheaper and increase the accuracy of the census-taking process.

The Census task is to obtain the following:

- Name
- Gender
- Marital Status
- Date of Birth
- Hispanic Origin
- Race

This task is complicated because of the nature of telephone speech, and the problems of speaker variability, background noise, and callers giving creative, unexpected or unresponsive answers.

37.3.2 Stages of the Project

The project started with designing and testing data collection protocols. Then data was taken and laboriously transcribed. This training data was used to develop a spoken language system. The system was then evaluated and refined.

Protocol design involves determining how to best ask a question; it must be easy to use and produce only the desired responses. CSLU has found that people don't usually think about how to ask a question. Structuring the problem is important to successful census information gathering.

An example of a bad prompt is:

- "What is your marital status?"

To which people have responded

- "Unhappy"
- "On hold"

Some examples of good prompts are:

- "Have you ever been married? Please say yes or no."

If *yes* is recognized, then prompt

- "Are you now married, widowed, divorced or separated?"

The next stage of the project was to collect data with the protocol mentioned above. Callers were solicited by the Census from 12 geographical regions. The callers called a toll-free number and answered both census questions and evaluation questions. Speech recognition was used to complete the questionnaire. CSLU received more than 4,000 complete calls. 1200 calls were laboriously hand-labelled and marked with a Behavioral Code. This is the basis for a speech corpus that will soon be released for use by other research groups.

An example set of behavioral codes (for the marital status question) is shown below:

Question– "Have you ever been married?"

- AA1 'yes'
- AA2 'yes I am'
- AA3 'unfortunately'
- QA 'well, sort of'
- IA1 'have you?'
- IA2 [says nothing]
- IA3 [garbled speech]

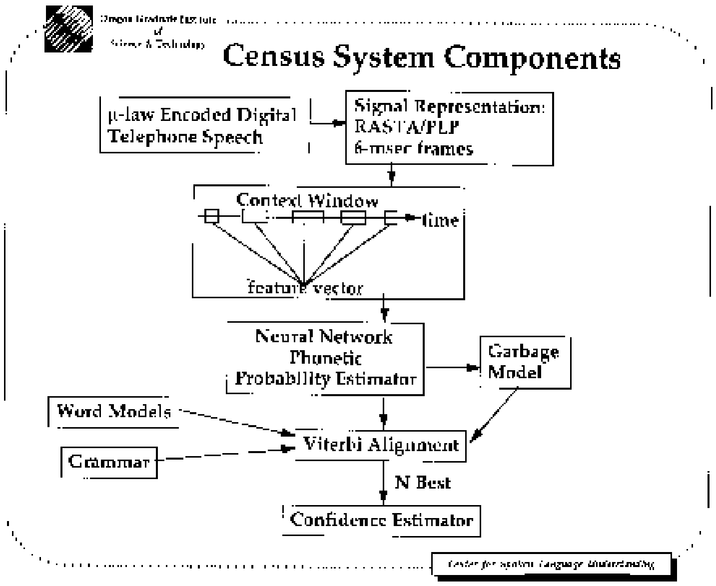

Figure 37.3: Census Recognition System Components

- RC 'does common law marriage count?'
- IN '...s' (interrupts)
- DK 'I'm not sure'
- RF 'That's none of your business.'
- O 'What was the question again?'

The responses are classified as informative or not informative, and it turns out that most people, 99.5%, give informative responses, though they may not answer the question in the manner the designer of the protocol intended. The general public is reasonably good at answering a question properly, if the question is asked with the right constraints.

The recognition system, based on the hybrid MLP-HMM concept seen in previous lectures, performs a PLP and voicing analysis on the speech input. Then task-specific, frame-based phonetic classification is done and a Viterbi search is used to do key-word spotting on the output. The 2-best answers are compared as a measure of confidence and the answer is output.

The census recognition system has many interacting parts (Figure 37.3). The system uses the garbage-word model idea developed by Hervé Bourlard. Garbage words are intended to match extraneous speech. For each time frame, the top n phones are averaged to produce a probability for a "garbage-phone". The word spotting system looks for a target word in the midst of silence, hissing or extraneous ("garbage") speech.

Context-dependent word modelling is used. Each phone has a word-dependant context associated with it (Figure 37.4). For a limited vocabulary task, a separate classifier can be trained for each word. However, if the specification of the problem then changes, as it often did for the Census project, it is more difficult to retarget this recognition system. Task independent word modeling uses broad categories for context instead of right and left phones, and this system can more easily be changed for different requirements.

The recognition system performs well; the results shown below, in Figure 37.5.

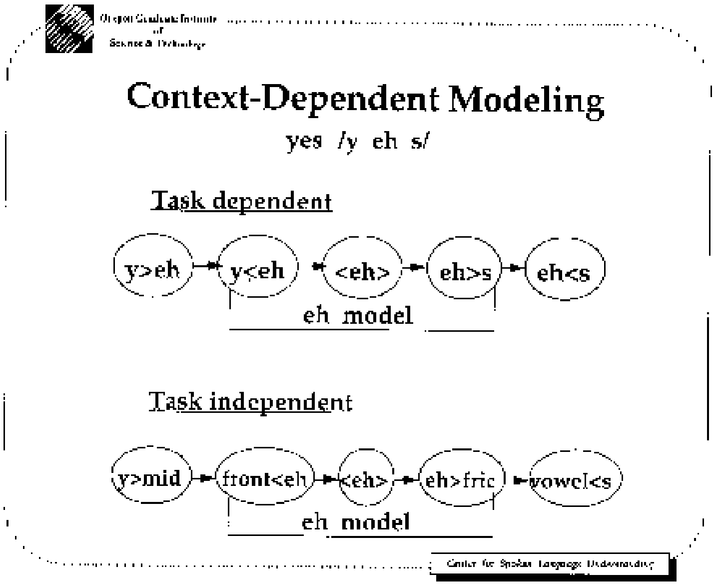

Figure 37.4: Context-dependent Modeling

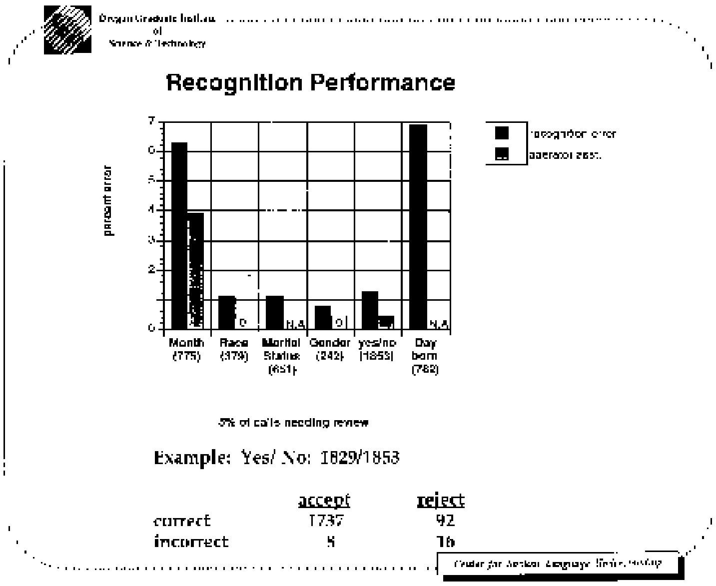

Figure 37.5: Recognition Performance

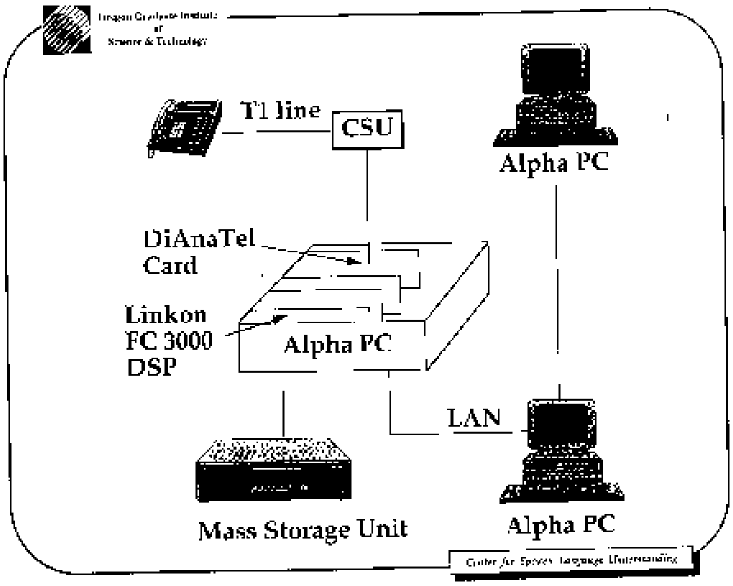

Figure 37.6: Census Recognition System Hardware

A confidence threshold was set so that if the system was not sure of an answer, it would reject the utterance. This reduced the errors by about half. However, this bring up the problem of what to do when an utterance is rejected; should the system bail out to another protocol, a human operator or what?

The phase 2 prototype system was designed to handle 24 simultaneous calls, both Spanish and English, incorporate an accurate, robust, graceful dialogue and be self-evaluating and allow for hand-off of the caller to an operator. For this last capability, the system builders had to reinvent handling "barge-in" where the caller speaks over the prompt, perhaps asking for the operator.

37.3.3 Results

The project now became a big system building effort, with lots of implementation details. It was eventually deployed in Jefferson, Indiana.

The finished system included six DEC Alpha PC's and could handle eight calls simultaneously (Figure 37.6). As it turned out, the system never had to handle as many as eight calls at once.

When the system was actually deployed, it was deeply buried in the Census menu of choices, so very few people actually utilized it. The operators all liked the easy-to-use interface, which automatically highlighted for the operator the progress of the call and what information was left to be filled in (Figure 37.7). The operators could later listen to the calls and correct information. When the caller bailed-out to talking to the operator, the operator was able to pick up where the caller left off.

The system was also stress-tested recently at CSLU and it seemed to work well. The results from this test are now being analyzed.

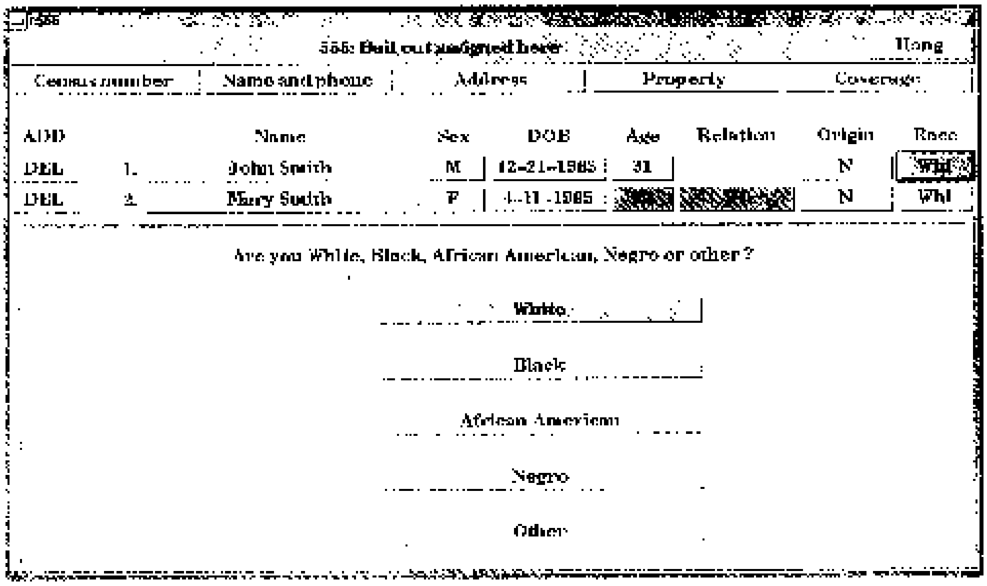

Figure 37.7: Census Recognition System Operator User Interface

37.3.4 What is the Future?

CSLU sees research issues everywhere. In the short term, CSLU is studying the difficult problem of rapid prototyping; how to retarget recognition systems quickly. Also, CSLU is studying dialog research, a part of speech recognition systems that is not often concentrated on.